Searching the Unsearchable: Inducing Serendipitous Insights

José Campos[1], A. Dias de Figueiredo[2]

[1] Departamento de Informática, Escola Superior de Tecnologia de Viseu,
Campus Politécnico de Repeses, 3504 -510 Viseu, Portugal
jcampos@di.estv.ipv.pt
[2] Departamento de Engenharia Informática, DEI/CISUC, Universidade de Coimbra,
Pinhal de Marrocos, 3030 Coimbra, Portugal
adf@dei.uc.pt

Abstract

Although no serious efforts seem to have been devoted yet to the theoretical and experimental study of the phenomenon, the web is recognizably a well suited medium for information encountering, the accidental discovery of information that is not sought for. This is the very essence of serendipity, the faculty of making fortunate and unexpected discoveries by accident. This paper presents Max, a software agent that uses simple information retrieval techniques and heuristic search to wander on the Internet and uncover useful, and not sought for, information that may stimulate serendipitous insights.

Introduction

Information retrieval usually assumes that the users know what they are searching for. Although this happens most of the time, namely when search engines are used, wandering on large information spaces, like the web, sometimes takes place with no specified goals. The web is, in fact, recognized as a well suited medium for information encountering (Erdelez 1996a; Erdelez 1996b; Toms 1996), the accidental discovery of information not sought for, and wandering on the web seems to be a quite usual user behavior leading to such serendipitous discoveries (Lieberman 1995; Toms 1996; Rosenfeld and Morville 1999).

Toms describes three typical ways in which people acquire information (Toms 2000):

- seeking information about a well-defined object;
- seeking information about an object that cannot be fully described, but will be recognized on sight; and,
- acquiring information in an accidental, incidental, or serendipitous manner.

The focus of our study is on the last topic. We believe that it is possible to induce and facilitate serendipity through the use of a special-purpose designed system. Though we agree with van Andel and Bourcier, that it is impossible to program serendipity (van Andel and Bourcier 1995), our key concern is that of programming *for* serendipity.

The design and development of such a system – that we have called Max – was based on research on the nature of insight, serendipity and creativity, inspired by the proposals of a number of authors, whose contributions we briefly comment.

Edward de Bono contributed to our view with the distinction between vertical and lateral thinking (de Bono 1990). While vertical thinking is selective and sequential, lateral thinking is generative and can make jumps. While vertical thinking concentrates on what are supposed to be the relevant aspects and excludes the irrelevant ones, lateral thinking welcomes any accidental and not sought for event. When a user is searching the web for some well defined object, vertical thinking is being done. Conversely, lateral thinking is the primary mental behavior when pure browsing activities are carried out. Within this framework, lateral thinking is likely to help developing an awareness of serendipitous events.

Thus, we have decided that Max should be programmed to present the user with information that follows the principles of lateral thinking: not excluding on the basis of immediate relevance, helping in delaying critical judgments, providing new entry points, etc.

In addition to those formulations, de Bono also proposed some practical techniques for lateral thinking that can be, to some extent, easily coded in a computer program, such as random stimulation, fractionation, the use of analogies, the selection of entry points, the reversal method, etc.

The concept of vertical thinking is deeply related to Kuhn's concept of normal science – characterized by the sequentially directed behavior of the scientist that attempts to articulate and extend an existing paradigm (Kuhn 1996). Paradigm shifts often occur by serendipity, when some unexpected event cannot be explained by the accepted paradigm, leading to accidental discoveries and sometimes to scientific revolutions (Kuhn 1996; Roberts 1989).

The primary goal of Max is that of stimulating the user with the precise information needed to provoke an insight or, in extreme cases, a paradigm shift.

Roberts points out that serendipitous discoverers share dominant characteristics, such as sagacity, perception (also described as awareness), curiosity, flexible thinking (similar to de Bono's lateral thinking) and intensive preparation (Roberts 1989). This is an important conclusion, since the success of Max's suggestions depends at least as much on the user as on the performance of Max.

Csikszentmihaliy and Sawyer stress that insights tend to occur during ‘idle times”, after a period of incubation and preparation (Csikszentmihalyi and Saywer 1996). They also established that creative insights typically involve the integration of perspectives from more than one domain of knowledge, as also pointed out by Kuhn in respect to scientific revolutions (Kuhn 1996) and by de Bono regarding creative thinking (de Bono 1990).

The perspective that Csikszentmihaliy and Sawyer offer about insights, namely the role of "idle times" and of cross-domain integration, inspired our development of Max, which has been created as a means to explore the browsing behavior of a user (an "idle time" activity) by generating a cross-domain integration of the interest profiles of that user.

Implementation

Max is an agent that browses the web, in a simulation of the typical human browsing behavior, searching for information that may interest the user, specially information that the user is not focused upon. By offering such information, Max attempts to stimulate the creativity of the user by providing new entry points based on de Bono's proposed techniques and, hopefully, induce serendipitous insights.

The major pillars of Max's implementation are the use of information retrieval techniques and the heavy use of heuristic search in information spaces.

To simplify user interaction, all the exchanges with Max are made through e-mail. This not only saves on design and implementation, but actually offers a more natural way of communicating with a software agent, by increasing the sensation of talking with a rational and anthropomorphic entity, though no effort as been made on producing a natural language interface.

System's Architecture

Max is composed of two functionally independent modules: the learning module and the suggestions formulation module (Fig. 1).

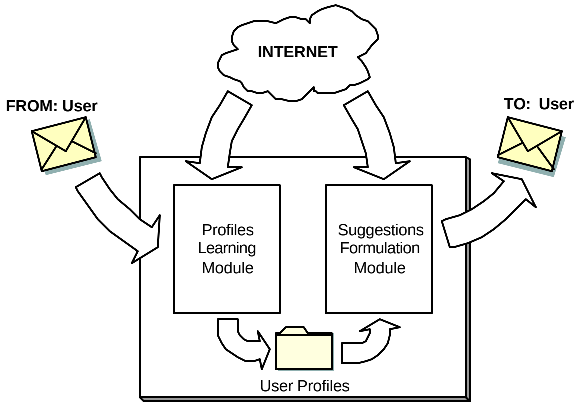

Figure 1: Max's architecture

User Profile Generation. Although profile generation was not a major issue in this project, some effort has been taken to integrate the structure of profiles with the information retrieval techniques used by the system. The goal of the profile generation process was not to generate automatically the profiles, on behalf of the user, by means of machine learning algorithms. The aim was to use those profiles whatever the generation process. Therefore, we assumed that the profiles were to be directly fed by the user through plain texts and the URLs of pages of interest to him. Since communication is by email, we have specified that the subject of the message should be used to label the "domain of interest" of the information sent to Max.

Figure 2 shows the architecture of Max's profile generation engine.

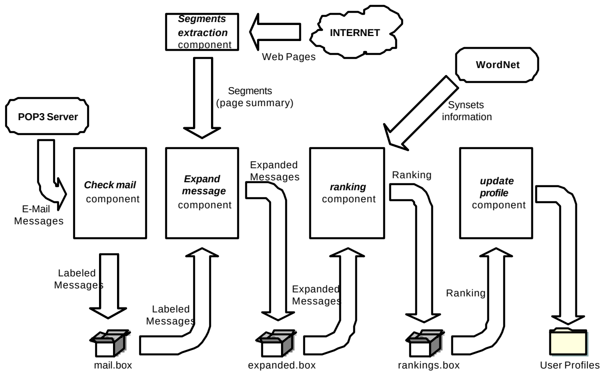

Figure 2: Profiles Learning Module

A first component is launched periodically to check e-mail from registered users. The body of the message is then passed on to the next component, with the identification of the user and the category label extracted from the subject field of the message.

The second component is in charge of the expansion of the URLs that may be embedded in the message text (a task that includes filtering HTML tags and traversing the page links recursively) so that the resulting data is just plain text. This task is simplified by resorting to the "segments extraction component" that splits the visited web page in text segments and returns the most relevant ones, providing a means of summarizing the contents of lengthy pages (Embley, Jiang and Ng 1999; Salton et al. 1996; Singhal and Salton 1995). An additional step is taken to eliminate stop-words (non-informative words, such as articles).

The third component of the system uses a *tf-ifd* measure – "term frequency-inverse document frequency", one of the simplest measurement methods existing in the information retrieval literature (Faloustos and Oard 1995; Salton and Buckley 1987) – to rank the concepts by their relevance to the user's message characterization. It is worth noting that the information retrieval methods we used works on concepts – WordNet synsets (Miller et al. 1993), not stems, the typical unit of information. By using WordNet synsets, we expect to increase the accuracy of profile generation by dealing more robustly with synonyms and compound words (ex. "artificial intelligence" and "social security"), yet supporting all the benefits of using stems. Another advantage of using concepts instead of stems is that this approach brings the problem to a higher level of abstraction, closely related to our purposes.

Finally, the fourth component simply merges the ranked data with the existing profile, following what we call the “learning parameters” (Sheth 1994) (the “learning parameters” let us tune the agent to be more or less conservative to new input information) and maintaining the normalization of the concepts' weights.

The Wandering Process. The search of interesting information is made by launching a *Google*$^{\text{TM}}$ query with some specially chosen words. The resulting URLs are browsed in a best-first style (Russel and Norvig 1995), heuristically directed by the user’s interest profiles.

Figure 3 shows the architecture of Max’s wandering process.

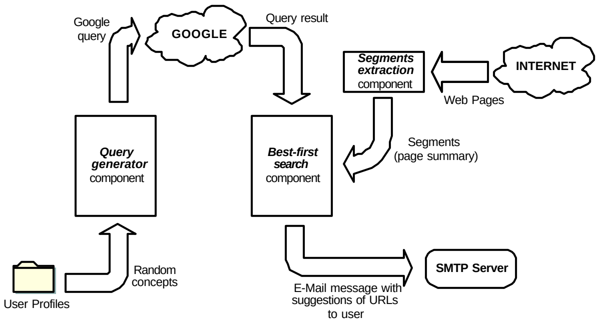
Google query
GOOGLE
Query result
Segments extraction component
INTERNET
Web Pages
Query generator component
Best-first search component
Segments (page summary)
SMTP Server
Random concepts
User Profiles
E-Mail message with suggestions of URLs to user

Figure 3: Suggestions formulation module.

The first component is responsible to generate a proper *Google*$^{\text{TM}}$ query. Following the suggestion of de Bono of using random stimulation and analogies (de Bono 1990), some domain profiles are chosen randomly to be the source profiles. From those profiles, some words are selected randomly, following an exponential probabilistic distribution (which means that we concentrate on the most relevant synsets, though not discarding the least relevant ones). Some of those synsets are used laterally: Max retrieves their coordinates[1] instead of using them directly. The selected words are merged in an unique query. This is our first attempt to perform the cross-domain integration suggested by Csikszentmihalyi and Sawyer (Csikszentmihalyi and Sawyer 1996).

Given that the URLs returned by *Google*$^{\text{TM}}$ are web pages with, typically, several links, and assuming that the linked pages are semantically related to each other in some degree, we implemented the wandering of Max using a best-first search, guided by a heuristic function. The wandering is quality and time limited by thresholds. When the search is over, the best-ever visited page address (from the point of view of the heuristic function) is sent to the user by e-mail.

Max’s Knowledge

As mentioned earlier, all the operations of Max are based on WordNet and WordNet’s synsets (Miller et al. 1993). WordNet is a huge lexical database for English whose design was inspired by psycholinguistic theories of human lexical memory. Nouns, verbs and adjectives are stored in synsets (synonym sets). A synset represents a lexical concept, which is stored along with a set of underlying relations to other concepts, thus forming a conceptual map.

WordNet is present in every phase of Max’s operation. During the learning process, Max transforms the text in a set of concepts, thought no relations are explicitly established between them. The knowledge that Max holds about the user (the user profiles) is a database of concepts. When a wandering process is started, Max chooses some concepts from the profiles and manipulates some of them before putting up the search engine query. The information retrieval methodology was also converted to work on concepts: the vectorization of the web pages uses concepts for each dimension of the vectors.

The use of synsets brings many direct benefits:

- abstraction is raised to a more appropriate level: we are not dealing with words anymore, but with concepts;
- synonyms are treated as being the same concept: this cannot be handled by stems – the usual atomic unit of information; this weakness of the stems leads to an underestimation of the weights of the concepts;
- compound words are handled properly: for instance, “artificial intelligence”, although being a single, unique concept, is usually decomposed into two distinct stems, “artific” and “intellig”; this, too, leads to an underestimation of the weights of the concepts;
- concepts manipulation: since WordNet establishes relations between concepts, it is easier to perform “lateral thinking experiments”; for instance, it is possible to obtain synonyms, antonyms, generalizations, specializations, coordinates, to compute concept distances, to generate analogies, and so on.

The most relevant drawback of using synsets instead of the usual stems is the overload introduced in the information retrieval methods, namely the vector transformation of the visited web pages. Now, instead of reducing a word to its stem (which is a fast and lightweight operation), the information retrieval methods need to seek the WordNet database for candidate concepts, follow disambiguation procedures, and handle structured data types that contain the concept information.

Nonetheless, this disadvantage is somehow diminished by the strong benefits WordNet provides.

The Heuristic Evaluation

The goal of the heuristic function is to guide Max during his wandering. It is out of doubt that such a heuristic function is very difficult to define. In the ideal world, the heuristic function would have to rate with maximum value the pages that would bring to the user the precise information needed to trigger a serendipitous insight in his mind.

Our first attempt to define the heuristic function has been based on the assumption that cross-domain integration is a strong and valid heuristic that may lead the wandering somewhere in the

[1] Two coordinate terms have the same hypernym. For instance, the concepts {discovery, breakthrough, find}, {revelation} and {flash} are coordinate terms, with hypermym {insight, brainstorm, brainwave} (Miller et al. 1993).

web, where several concepts from different knowledge domains may be present.

The heuristic function is solely used by the best-first search to sort the children pages. The most promising page is then explored.

Since we cannot set the heuristic value of a page before analyzing it (i.e., before extracting and weighting its concepts), we had to transform the best-first search to support this reversed feature. In the typical best-first algorithm, children nodes are expanded if they are promising. In our approach, a child page has to be expanded (visited) to assess if it is promising.

To accelerate page analysis, which is a heavy task, we extract from each page the central segment – a heuristically selected subset of the integral text that summarizes the whole page, as done in the learning process (Embley, Jiang and Ng 1999) – and use it for heuristic analysis.

The value of the heuristic evaluation is computed using the external product of the vector-based representation of the page and the profiles (Salton and Buckley 1987):

$$\vec{p} \cdot \vec{w} = \|\vec{p}\| \, \|\vec{w}\| \cos \alpha = \cos \alpha$$

where $\vec{p}$ is the profiles vector, $\vec{w}$ is the web page vector and α is the angle between the two vectors. The vector representing the page is calculated through *tf-idf*. Given that the external product of two vectors is equal to the cosine of α (if the vectors are normalized), the page and profiles are more similar as the angle approaches zero.

It is important to notice that the intended purpose of the heuristic function is not to find web pages directly related to the interests of the user, as many other systems do (Lieberman 1995; Pazzanni et al. 1996; Edwards et al. 1996; Moukas 1996). Instead, since we use many profiles simultaneously for evaluation, the cross-domain behavior of the heuristic is expected to guide page selection towards not so related pieces of information.

Types of Expected Results

The mission of Max is to stimulate the user with unexpected information. Not any kind of information will trigger the creativity of the user. As pointed out by van Andel and Roberts, there are essentially three personal characteristics of serendipitous discoverers: sagacity, adequate preparation, and curiosity (Roberts 1989; van Andel and Bourcier 1995).

Sagacity is the characteristic that catches the attention of the user. Sagacity is deeply influenced by the background knowledge of the user. After sagacity has expressed itself, there are two possible paths: if the preparation of the user matches the input, an insight may occur. Otherwise, curiosity may lead to dedicated investigation.

Usually, the user expectations lie on the accuracy of Max's suggestions, but Max only knows a small and inaccurate portion of the user's real interests and knowledge.

This leads us to postulate six possible categories in the value of Max's suggestions:

- **category 1:** the suggested pages were already known – the suggestion has no value at all;
- **category 2:** the suggested pages were not known and did not belong to any domain of interest – the suggestion has little value. Although not interesting at the present moment, the suggested pages may have some usefulness in the future (in the context of lateral thinking);
- **category 3:** the suggested pages were not known, but belong to some domain of interest – the suggestion has little value, since the user could have reached those pages otherwise (ex.: using a search engine);
- **category 4:** the suggested pages were not expected, but they are slightly related to some domain of interest – the suggestion is valuable, and it could hardly have been found by the user;
- **category 5:** the suggested pages were not expected, they did not belong to any current domain of interest of the user, but they sparked in him a new interest – the suggestion is extremely valuable since it is very improbable that the user would **ever** find the page on himself;
- **category 6:** the suggested pages establish a new connection between two current domains of interest – the suggestion is extremely valuable (an insight occurred);

From the beginning of the development of Max, we did not expect overwhelming results in a short time. Our main ambition was to obtain *some* kind of interesting input from Max, specially within categories 3 and 4, that could guide the fine tuning of further implementation decisions. It is evident that, though the performance of Max is very important, the sagacity, preparation and curiosity of the user are crucial. So, albeit our most optimistic hope, the expected results were to be essentially subjective.

Though within such a subjective frame of mind, we are glad to notice that, much above our expectations, Max has been offering us a number of quite promising suggestions.

Empirical Results

At the time this paper was written, we had been using Max for two months only, which is recognizably insufficient to establish any accurate conclusions (even assuming the subjective nature of the results).

From around 100 messages Max sent with suggestions, two were found to belong to category 5 and five to category 4, a result that has surprised us. While these results may seem poor, we did consider them, in fact, quite productive:

- the suggestions from category 4 (pages slightly related with the domains of interests) were considered very valuable. Those pages and subjects were unknown but had something in common with the domains of interests. They could hardly be found without Max. Yet, after their presentation, they brought new directions and perspectives;
- the suggestions from category 5 (pages that did not belong to any domain of interest but sparked new

interests) led to a *sui generis* reaction: one was considered very important, but the other not so valuable, thought interesting;

Some conclusions can be drawn from these results:

- although the number of valuable suggestions may seem quite low, the benefits of using Max defeated the penalties. If those pages had not been suggested, they would quite likely never been known by the user;
- the suggestions that fell into category 4 were considered very difficult to find without Max, even if resorting to a search engine. The reason pointed out was that the subjects of those pages slightly overlapped the domains of interest. A quite intensive effort of divergence from those domains of interest would have been necessary to reach the suggested pages: first, to build a query string for the search engine; then, to guide the browsing towards the pages suggested by Max;
- Even when some pages are considered interesting, this does not mean that they are valuable. This situation happened with one suggestion of category 5, which confirms our feeling about the deep subjectivity of results in this field.

Future Work

The results from those first experiments with Max is convincing us that this approach to "programming *for* serendipity" may be very promising in the near future. With this in mind, we are planning to improve Max in several ways:

- Improving its capacity to disambiguate the concepts during the concept extraction process;
- Resorting to more powerful mechanisms for the generation of divergence, namely through the use of metaphors;
- Improving the heuristic function, namely by extracting different parts of the web page, to be analyzed *per se* (e.g. page title, formatting emphasis, links, etc.);
- Extending the usage of WordNet, namely, by exploring better the relations between concepts; and,
- Trying to raise the abstraction level to a higher stage, for instance, that of ideas.

Conclusions

It is acknowledged that the web is a well suited medium for accidental information discovery. Although not as a primary information behavior, browsing and accidental, serendipitous, discovery of information are tightly related.

This paper presents an ongoing project that attempts to cast some light on the possibility of inducing serendipity through the use of specially designed systems.

We have presented Max, a software agent that mimics the browsing behavior of users navigating the web just for the sake of wandering. Max has some knowledge about the user interests. While wandering, Max stays aware of possible interesting pages ("interesting" in our context means "potentially insightful").

The user interests are coded in profiles, one profile per domain of interest. The profiles are directly fed by the user, who sends to Max e-mail messages that contain plain text and URLs.

The Max wandering process begins with a *Google*^{TM} search of randomly chosen words from the profiles. To guide Max, a best-first search is performed over the resulting pages. This search is guided by a heuristic function that gives more weight to pages that have more cross-domain integration.

Even though we are very hopeful about the conclusions of the project, it is extremely difficult to measure its success due to the subjective nature of the effect the system has on the users.

Nevertheless, Max has suggested some pages that fall in categories 4 (pages slightly related to the domains of interest) and 5 (pages that sparked new interests). No suggestions fall in the category 6 yet. This results had encouraged us to proceed with further investigation and developments, namely, the use of metaphors, the tuning of concept disambiguation, the improvement of the heuristic function and extending the usage of WordNet.

Acknowledgments

This work is partly supported by the POSI Program of the Portuguese Foundation for Science and Technology and European Union FEDER. The first author acknowledges the support received from the Computer Science Department of Escola Superior de Tecnologia de Viseu.

References

Csikszentmihalyi, M.; Sawyer, K. 1996. *Creative Insight: The Social Dimension of a Solitary Moment*. The Nature of Insight. MIT Press.

De Bono. E.1990. *Lateral Thinking*. Penguin Books.

Edwards, P.; Bayer, D.; Green, C. L.; and Payne, T.R. 1996. Experience with Learning Agents which Manage Internet-Based Information. In *AAAI Spring Symposium on Machine Learning in Information Access*, 31-40. Menlo Park, CA: AAAI Press.

Embley, D. W.; Jiang, Y.; and Ng, Y. K. 1999. *Record-Boundary Discovery in Web Documents*. In Proceedings of the 1999 ACM SIGMOD International Conference on Management of Data, Record. 28(2), pages 467-477. Philadelphia, Pennsylvania: Associations for Computing Machinery.

Erdelez, S. 1996a. *Information Encountering: A Conceptual Framework for Accidental Information Discovery*. In Pertti Vakkari et al. (Eds.), Information Seeking in Context. Proceedings of an International Conference on Research in Information Needs, Seeking and Use in Different Contexts. Tampere, Finland: Taylor Graham.

Erdelez, Sanda. 1996b. *Information Encountering on the Internet*. In M. Williams (Ed.), Proceedings of the 17th National Online Meeting. New York. Medford, NJ: Information Today.

Faloutsos, C.; and Oard, D. 1995. A Survey of Information Retrieval and Filtering Methods, Technical Report CS-TR-3514, Dept. of Computer Science, Univ. of Maryland.

Kuhn, T. S. 1996. The Structure of Scientific Revolutions, Third Edition. The University of Chicago Press.

Lieberman, H. 1995. Letizia: An Agent That Assists Web Browsing. In *Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence*, Morgan Kaufmann, 924-929. Montreal, Quebec, Canada: AAAI Press.

Miller, G. A.; Beckwith, R.; Fellbaum, C.; Gross, D.; and Miller, K. 1990. Introduction to WordNet: An On-line Lexical Database. *International Journal of Lexicography*, vol. 3(4), 235-244.

Moukas, A. 1996. Amalthaea: Information Discovery and Filtering using a Multiagent Evolving Ecosystem. In Proceedings of the First International Conference on the Practical Application of Intelligent Agents and Multiagents Technology, London, UK.

Nielsen, J. 1999. *Designing Web Usability*. New Riders Publishing.

Pazzani, M.; Muramatsu, J. ; and Billus D. 1996. Syskill & Webert: Indentifying Interesting Web Sites. In *Proceedings of the Thirteen Nacional Conference on Artificial Intelligence*, 54-61. Portland, Oregon: AAAI Press.

Rosenfeld, L.; and Morville, P. 1998. *Information Architecture for the World Wide Web*. O'Reilly & Associates.

Roberts, R. M. 1989. *Serendipity: Accidental Discoveries in Science*. John Wiley & Sons, Inc.

Russel, S.; and Norvig, P. 1995. *Artificial Intelligence: A Modern Approach*, 92-121. Prentice Hall International Editions.

Salton, G.; and Burckey, C. 1987. Term Weighting Approaches in Automatic Text Retrieval, Technical Report 87-881, Department of Computer Science, Cornell University.

Salton, G.; Singhal, A.; Buckley, C; and Mitra, M. 1996. Automatic Text Decomposition Using Text Segments and Text Themes. In Proceedings of the Seventh ACM Conference on Hypertext, Washington DC, USA.: Associations for Computing Machinery.

Sheth. B. 1994. A Learning Approach to Personalized Information Filtering. Master's Thesis, Dept. of Electrical Engineering and Computer Science, MIT.

Singhal, A.; and Salton, G. 1995. Automatic Text Browsing Using Vector Space Model. In Proceedings of the Fifth Dual-Use Technologies and Applications Conference, 318-324, Utica/Rome, NY.

Toms, E. 1996 Information Exploration of the Third Kind: The Concept of Chance Encounters. In Proceeding of the CHI98 Workshop on Information Exploration, Los Angeles: Associations for Computing Machinery.

Toms, E. 2000. Serendipitous Information Retrieval. In Proceedings of the First DELOS Network of Excellence Workshop on Information Seeking, Searching and Querying in Digital Libraries, Zurich, Switzerland: European Research Consortium for Informatics and Mathematics.

Van Andel, P; and Bourcier, D. 1990. *Peut-on Programmer la Sérendipité? L'Ordinateur, le Droit et l'Interpretation de l'inattendu*. Netherlands Institute for Advanced Studies, Royal Netherlands Academy of Arts and Science, Wassenaar, the Netherlands.